A Lightweight Empirical Analysis of Similarity-Based Membership Inference Attacks Against Retrieval-Augmented Generation Systems

Aqsa Shabbir
Department of Computer Engineering
Bilkent University, Ankara, Turkey
Email: aqsa.shabbir@bilkent.edu.tr

***Abstract*—Retrieval-Augmented Generation (RAG) systems integrate an external knowledge base with a large language model (LLM), improving factual grounding while introducing novel privacy risks. Previous work has shown that RAG systems may leak membership information about documents stored in the retrieval corpus. In this study, we empirically evaluate the simplest form of membership inference attack—Similarity-Based MIA (S-MIA)—on a small and computationally manageable RAG pipeline designed to run in constrained environments such as Google Colab. Using the AgNews dataset, a Flan-T5-small model for generation, and MiniLM embeddings for retrieval and similarity measurement, we demonstrate that even lightweight RAG systems exhibit measurable membership leakage. Across three experimental seeds, S-MIA achieves an average AUC of 0.633 and an average TPR of 21.33% at a 5% false-positive rate. While leakage is moderate compared to large-model settings, the results validate the vulnerability of RAG architectures even under minimal-resource configurations. The findings also serve as a foundation for future extensions toward implementing the full Difficulty-Calibrated MIA (DC-MIA) pipeline.**

***Index Terms*—Retrieval-Augmented Generation, Membership Inference, Privacy, Large Language Models, Similarity Attacks.**

I. Introduction

Large Language Models (LLMs) have become a central component in modern natural language systems, but their tendency to memorize training data raises significant privacy concerns. Retrieval-Augmented Generation (RAG) architectures extend LLMs by querying an external corpus using dense embeddings, enabling the model to incorporate retrieved passages into its responses. While this improves factual accuracy, it also creates a new attack surface: an adversary may exploit the interaction between retriever and generator to infer whether a specific document exists in the RAG knowledge base.

Recent research, including the RAG-Leaks attack framework, demonstrates that RAG systems are susceptible to membership inference attacks (MIA). In particular, observing the semantic similarity between model-generated answers and ground-truth references can reveal latent correlations between retrieval behavior and stored corpus entries. However, the majority of existing work assumes large LLMs (e.g., Mistral-7B, Llama-3) and substantial retrieval corpora, which imposes computational challenges for small-scale environments such as educational or exploratory settings.

The goal of this report is to provide a detailed and reproducible study of S-MIA on a lightweight RAG deployment. Our motivation is twofold: (i) to verify whether measurable leakage exists even under resource constraints, and (ii) to establish an experimental scaffold that can later be extended to full DC-MIA implementations. Despite using a small LLM and a reduced corpus, our results show consistent membership signals across random seeds.

II. Background

A. Retrieval-Augmented Generation (RAG)

A RAG system consists of two main components: a retriever and a generator. The retriever encodes the query into an embedding space and retrieves semantically relevant documents from a knowledge base. These documents are appended as context to the LLM, which then generates an answer conditioned on both the query and retrieved content. This mechanism increases factual grounding but also induces privacy leakage, because retrieved content may strongly influence the generator's outputs.

B. Membership Inference Attacks (MIA)

Membership inference seeks to determine whether a target sample belongs to the dataset used by the system. In RAG settings, the question becomes: *Was a specific document stored in the retrieval corpus?* Similarity-Based MIA (S-MIA) infers membership by comparing the cosine similarity between the generated response and the ground-truth answer. If the system retrieves the exact document (because it is in the KB), the generated answer tends to be semantically closer to the true reference.

III. Methodology

A. RAG Construction

Our RAG pipeline uses:

- **Retriever:** MiniLM-L6-v2 embeddings (SentenceTransformers), encoded into a FAISS IndexFlatIP for fast cosine similarity search.
- **Generator:** Flan-T5-small, chosen for its lightweight inference footprint.
- **Knowledge Base:** 300 AgNews samples, each partitioned into a query (first half of text) and answer (second half).

During inference, for each query, the top-2 KB documents are retrieved and inserted into a prompt template. The LLM produces an answer, which is then embedded using MiniLM to compute similarity against the ground-truth answer.

B. Similarity-Based Membership Inference (S-MIA)

Given a sample (q, a), the attack computes:

$$S = \text{cosine}(\text{Emb}(\text{LLM}(q)), \text{Emb}(a)).$$

Higher similarity indicates that the RAG system produced an answer closer to the correct ground truth. A threshold τ is chosen using ROC curve analysis such that the false positive rate (FPR) does not exceed 5%. Samples with $S > \tau$ are predicted as members.

IV. Experimental Setup

A. Dataset Preparation

We use 3000 AgNews samples, from which:

- 300 documents populate the RAG knowledge base,
- 50 KB documents serve as member evaluation samples,
- 50 non-KB documents serve as non-member evaluation samples.

Each document is split into query and answer portions.

B. Hardware Environment

Experiments were performed on Google Colab CPU runtime. Due to limited memory, we intentionally use a small LLM and limit retrieval depth. Our objective is methodological validation rather than reproducing full RAG-Leaks attack strength.

C. Multi-Seed Evaluation

To ensure stability, we evaluate the attack across three random seeds: 42, 43, and 44. For each seed we re-shuffle the dataset, rebuild the KB, and recompute similarity scores.

V. Results

A. Overall Membership Leakage

Table I presents the aggregated results.

TABLE I
Multi-seed S-MIA performance summary.

Seed	AUC	TPR @ 5% FPR
42	0.604	20.0%
43	0.629	28.0%
44	0.667	16.0%
Mean	**0.633**	**21.33%**
Std	0.026	4.99%

Even in this restricted setup, the attack consistently outperforms random guessing (AUC=0.5). The average AUC of 0.633 indicates moderate separability between members and non-members.

B. Similarity Distribution Analysis

Members tend to produce higher cosine similarities due to direct retrieval from the KB, while non-members exhibit broader, lower-scoring distributions. However, certain non-members with semantically similar content achieve high similarities, creating overlap regions that limit attack accuracy.

Example: High-Similarity Member

Query: President OKs More Colombia Assistance (AP) President Bush said Tuesday the ...

Ground-Truth Answer: U.S. government will continue to assist Colombia in interdicting aircraft suspected of drug trafficking ...

Observation: Because this document is stored inside the RAG knowledge base, the retriever consistently surfaces it as top-ranked evidence. The Flan-T5-small generator then reproduces an answer that is semantically extremely close to the ground truth, resulting in a cosine similarity of **0.9819**. This qualifies as a clear membership indicator.

Example: High-Similarity Non-Member

Query: Last Call for Investors to Bid on Google Time is running out for prospective investors ...

Ground-Truth Answer: Investors must submit their offers to buy shares of Google Inc., the Web's leading search company ...

Observation: Although this sample is *not* present in the knowledge base, the retriever locates highly similar financial news articles. The generator produces answers with content that strongly overlaps the true reference. This yields a similarity of **0.9494**, demonstrating that S-MIA may incorrectly flag certain semantically aligned non-members as members.

C. Attack Performance

To better understand the behavior of the similarity distributions and the separability achieved by the S-MIA attack, we include ROC curves and similarity histograms for each evaluated seed. These plots illustrate (i) the relative position of the similarity threshold τ, (ii) the degree of overlap between member and non-member distributions, and (iii) the placement of the TPR operating point at FPR=0.05.

VI. Conclusion

This report demonstrates that even a minimal RAG system running on CPU with a small LLM exhibits measurable membership leakage when subjected to a similarity-based membership inference attack. While the leakage is significantly weaker than reported in full RAG-Leaks implementations, our

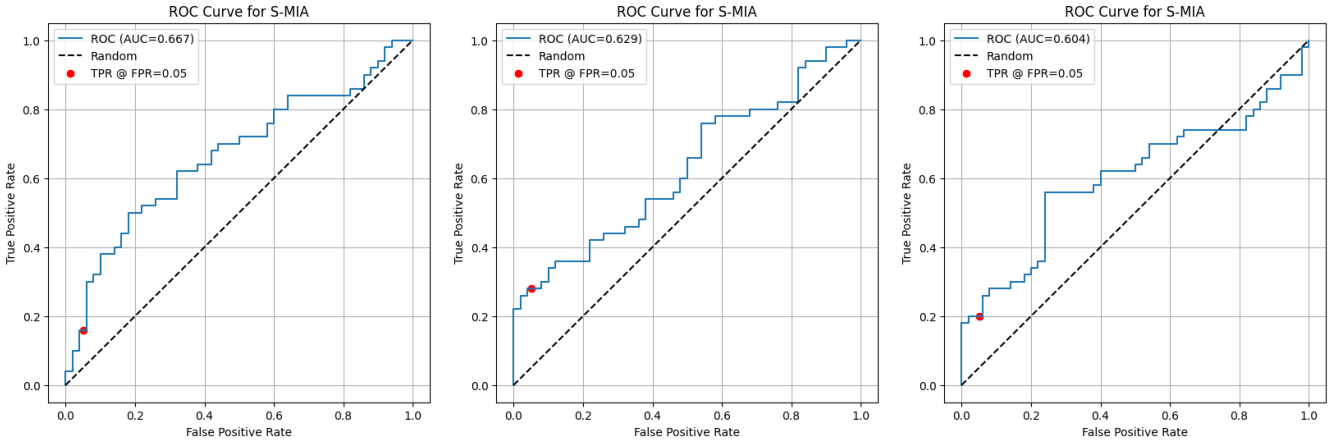

Fig. 1. ROC curves for S-MIA across three seeds. The AUC values vary between 0.604–0.667, indicating moderate separability. The red marker denotes the TPR achieved at the 5% FPR operating point.

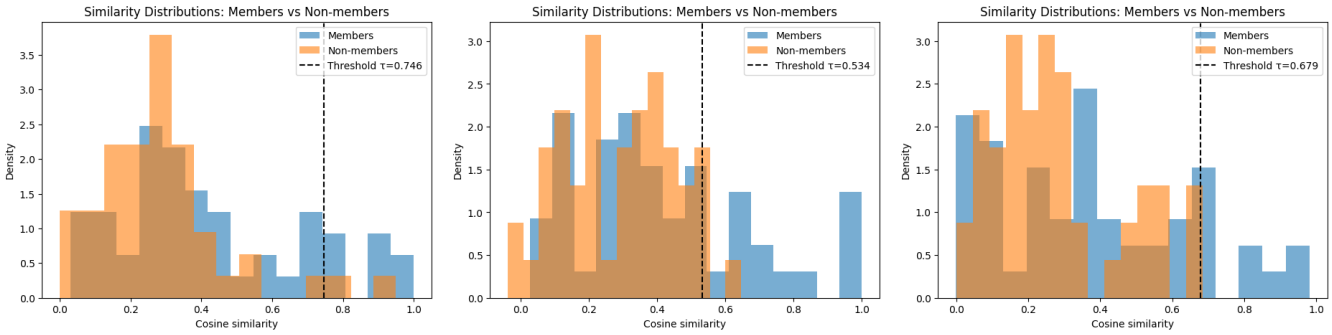

Fig. 2. Similarity score distributions for members vs. non-members across three seeds. Vertical dashed lines represent thresholds τ determined under an FPR constraint of 5%. Overlap in these distributions explains the moderate attack performance.

findings confirm that RAG architectures inherently encode membership signals due to the interaction between retriever and generator. The moderate AUC and low but non-trivial TPR at stringent FPR thresholds suggest that similarity alone is insufficient for strong inference, motivating the need for more sophisticated techniques such as Difficulty-Calibrated MIA (DC-MIA). Future work will extend this lightweight pipeline to include reference RAG construction, likelihood ratio modeling, and Gaussian-based calibration, bringing the experiment closer to the theoretical framework of the original RAG-Leaks attack.

REFERENCES

[1] X. Authors, "RAG-Leaks: Difficulty-Calibrated Membership Inference Attacks Against Retrieval-Augmented Generation Systems," 2024.
[2] W. Authors, "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression," in *NeurIPS*, 2020.
[3] C. Authors, "Flan-T5: Improved Instruction Fine-Tuning for Language Models," Google Research, 2022.